学校编码：10384　　分类号_____密级_____
学　　号：23020141153198　　UDC_____

厦门大学

硕士学位论文

基于稀疏表示的特征选择算法研究

Research of Feature Selection Algorithm Based on Sparse Representation

田毅炀

指导教师姓名：夏侯建兵 副教授
专业名称：计算机技术
论文提交日期：2017年4月
论文答辩时间：2017年5月
学位授予日期：2017年　月

指导教师：

答辩委员会主席：

2017年4月

厦门大学学位论文原创性声明

本人呈交的学位论文是本人在导师指导下，独立完成的研究成果。本人在论文写作中参考其他个人或集体已经发表的研究成果，均在文中以适当方式明确标明，并符合法律规范和《厦门大学研究生学术活动规范（试行）》。

另外，该学位论文为（　　）课题（组）的研究成果，获得（　　）课题（组）经费或实验室的资助，在（　　）实验室完成。（请在以上括号内填写课题或课题组负责人或实验室名称，未有此项声明内容的，可以不作特别声明。）

声明人（签名）：田毅炀

2017 年 5 月 11 日

厦门大学学位论文著作权使用声明

本人同意厦门大学根据《中华人民共和国学位条例暂行实施办法》等规定保留和使用此学位论文，并向主管部门或其指定机构送交学位论文（包括纸质版和电子版），允许学位论文进入厦门大学图书馆及其数据库被查阅、借阅。本人同意厦门大学将学位论文加入全国博士、硕士学位论文共建单位数据库进行检索，将学位论文的标题和摘要汇编出版，采用影印、缩印或者其它方式合理复制学位论文。

本学位论文属于：

（　　）1.经厦门大学保密委员会审查核定的保密学位论文，于　　年　　月　　日解密，解密后适用上述授权。

（ √ ）2.不保密，适用上述授权。

（请在以上相应括号内打“√”或填上相应内容。保密学位论文应是已经厦门大学保密委员会审定过的学位论文，未经厦门大学保密委员会审定的学位论文均为公开学位论文。此声明栏不填写的，默认为公开学位论文，均适用上述授权。）

声明人（签名）：田毅炀

2017 年 5 月 11 日

摘要

在模式识别学科中，特征选择作为其范畴内的一个重要方向，已经演变成近些年来的学习热点。在现实生活中，科学研究的成果已经渗透到很多行业，并在行业中获得实际应用。在学科研究和现实生活应用中，将会面对和处理庞大的数据。该数据往往样本数不多，但是其数据维数很大并且冗余特征多，对计算机的处理资源和处理实时性是很大的挑战，解决“维度灾难”的问题有非常重要的作用。所以特征选择作为数据处理的重要步骤，发挥关键的作用。

由于维度过大的原因，高维数据的回归问题是一个比较大的挑战，一个有效的解决方法就是特征选择。而基于稀疏表示的线性回归已经被证明在处理高维数据时非常有效。传统的稀疏表示的线性回归算法有 Lasso 算法，Lasso 算法通过最小化目标函数，用系数的绝对值作为压缩模型的系数，使得绝对值比较小的系数被压缩为 0，这样就可以去掉很多不重要的特征。由于 Lasso 在特征选择方法上的优点，得到了广泛的认可和使用。

为了解决高维数据所面临的特征选择问题，本文是在稀疏表示的线性回归模型上，进一步深入进行研究。结合 Lasso 线性回归模型，提出了具有辨别信息的特征选择模型，其特征变量与特征变量有很少的重复性，同时特征变量与响应变量存在很大的相关性。同时，基于 Lasso 模型，提出一种特征交互性特征选择方法。所选择特征体现了协变量和响应集变量高阶交互信息。本文在多个公开的数据集上进行测试。从实验测试结果中可以看出，提出的模型对于特征选择任务的分类准确性有了明显的提升。

关键词：特征选择；Lasso；交替方向乘子求解

Abstract

In the pattern recognition disciplines, the feature selection as an important direction within its scope, which has evolved into a hotspot in recent years. In real life, the results of scientific research have penetrated into many industries, and obtain practical application in the industries. In disciplinary research and real-life applications, we will face and deal with huge amounts of data. However, these data have a small number of samples, and its data dimension is large, at the same time, with same redundant features, which is a big challenge for the computer processing resources and processing real-time. To solve the "dimension of disaster" problem has a very important role. Therefore, feature selection plays an important role as an important step in data processing.

Because of the large dimension, the regression problem of high dimensional data is a relatively large challenge. An effective solution is the feature selection. While linear regression based on sparse representation has proven to be very effective in dealing with high dimensional data. The traditional sparse representation of the linear regression algorithm is Lasso algorithm. Through minimizing the objective function, with the absolute value of the coefficient as the compression model coefficients, making the absolute value of the smaller coefficients are compressed to 0, so Lasso algorithm can remove many unless features. Due to the advantages of Lasso in the feature selection method, it has been widely recognized and used.

In order to solve the feature selection problem faced by high-dimensional data, this paper is based on the sparse representation of the linear regression model, to do a further study. Based on the Lasso linear regression model, a feature selection model with discriminant information is proposed. The characteristic variables and feature variables have little repetition, and the characteristic variables have a great correlation with the response variables. At the same time, a feature selection method based on Lasso model is proposed. The selected feature reflects the higher order interaction information of the covariate and response variables. This article is tested on multiple open datasets. It can be seen from the experimental test results that the proposed model has improved the

classification accuracy of the feature selection task.

Keywords: Feature selection; Lasso; ADMM

目录

Contents

第一章 绪论

1.1 课题背景及研究意义

在科技持续发展的今天，人工智能等新兴科技也得到了进步。模式识别、机器学习作为人工智能技术的主要学科，已经得到广泛的重视，很多科技研究工作人员已经把机器学习等算法应用在实际场景中。该技术带来的应用在人们生活中的比例日益增长，为生活提供了更多的便捷，人们几乎天天都能看到关于智能技术研究所取得的新突破和新的实际应用。

同时，随着科技的发展，人们所获得的信息量越来越大。以网络信息为例，复杂而庞大的数据随着人们的生活和工作不断产生，在生活中的方方面面，各种信息数据都是以爆炸式的增长，比如社交网络中的图片、视频，日常超市购买，网上购物等。这些信息数据有共同一个特点，即它们都是行业中的重要的信息，需要即时对它们进行处理。尽管这些数据使得人们获得的信息更加充分，但是这些数据往往有很高的维度，而且数据量增长速度快，而且这些数据中往往存在很多的冗余特征，这对计算机处理和存储这些数据是非常大的挑战。这就是机器学习中“维度灾难”问题。解决这个问题，可以提高计算机的处理能力和存储性能，帮助人们获取更具有价值的数据信息。

在模式识别、机器学习领域中，为了解决大数据带来的“维度灾难”的问题，研究者们想出了很多办法，其中最主要的方法就是对数据进行降维。总体上来说，数据降维，就是在原始的空间维度中，将数据变换至另一个维度比较低的子空间。所处理的数据在这个空间中具有更低的数据维度，但仍保留具有对数据的表示性。一般来说，传统的数据降维，分别是特征提取和特征选择。特征提取是将原始数据转化成机器学习算法能识别的特征的过程。特征选择在原始的特征变量集合中选择好的特征变量子集合，是去掉无关的特征变量，达到保留相关性的特征变量的目的。在机器学习算法中，一般先对数据进行预处理，这就是进行特征选择的过程。经过特征选择的数据集再去训练学习分类器。特征选择过程需要使得重要的信息不会丢失，假如丢失了重要的信息则后续的学习过程将会失去好的性能。

其中存在的冗余特征是指对于学习过程的相关性小的特征，不同的学习过程要求特征选择步骤能够去掉这些特征。因此，特征选择作为模式识别、机器学习等人工智能领域的关键组成部分，一直是近几十年从业人员的主要研究方向。

专业研究人员针对这些问题，已经提出了很多特征选择方法，其中大概可以分成两类。一种是特征排名，一种是特征子集搜索。其中特征排名，主要是一个一个地去评价特征重要程度，根据评价结果对特征的重要程度进行排序，优先选择重要性大的特征。其代表方法有 FS(Fisher Score)[1]方法和 LS(Laplacian Score)[2]方法。而对于特征子集搜索方法，它根据某种准则来选择候选特征子集的重要性，并从中选择出最优的特征，其代表方法有 Lasso(Least Absolute Shrinkage and Selection Operator)[3]方法。

基于稀疏表示的特征选择已经得到深入的探索，其中 Lasso 方法已经被应用到很多问题中。然而 Lasso 方法仍然具有限制性，由于其通常考虑特征和类别的相互关联信息，从而忽略了特征与特征的相互关联信息。因此，造成有用信息丢失和冗余信息增加，从而影响的 Lasso 特征选择算法性能。所以，改进 Lasso 类型的特征选择算法，有非常重要的意义。

1.2 国内外研究现状

对于特征选择算法的研究起始于是上世纪 60 年代。最早的特征选择算法是在统计学信号处理领域，在那个年代信息量比较少，对于所处理的数据的规模比较小。进入 21 世纪，各个行业的应用逐渐增多，现存的特征选择算法已经无法满足日益增长的数据量要求。因此，更多的科学从业人员深入探索处理高维数据特征选择的新方法。

在上世纪 60 年代，Lewis 等人是最早开始发现和研究特征选择问题，当时主要的应用领域是在统计学信号处理，当时所用的实验数据是比较小的，相对的特征维度也比较低，特征与特征之间假设不会存在相互的关系。Krzanowski 在 1987 年提出 KP 算法，该算法在建立在 Procruste 分析的基础上，通过选择的特征子集，极大限度上保存数据中的多元结构信息。之后，Mao 提出了 Forward LSE 算法，对 PCA 的变换矩阵进行了向前搜索和向后搜索，通过这方法来选择特征，

但是这种搜索方法的时间复杂就被提高了很多[4]。

在 90 年代，新兴的研究方向让特征选择问题接受了新的挑战，如图像搜索、基因工程、文本分类的应用，使得所处理的数据开始变得巨大。而这些应用需要有高效的使用效率，更要求能够运用更多更大的数据量。这时候，特征选择问题被提到一个研究的新高度[5]。

在 1992 年，Kira 和 Rendell 等人提出了 Relief 算法[6]，Relief 算法通过剔除无关的特征，得到最终的特征子集。具体的做法为，Relief 算法通过一一计算特征变量与类别变量的相关程度，用来给每个特征定义权重大小。根据设定的阈值，将权重与阈值进行比较，保留大于该阈值的权重对应的特征，同时将小于该阈值所对应的特征给剔除掉，从而到达了特征选择的作用。但是 Relief 算法不能完全把冗余特征去除，往往只用于二类分类。1994 年 Almuallim 和 Dietterich，利用信息论来进行特征选择研究，提到选择最优的特征需要花费很大的搜索空间[7]。同年，Kononenko 等人对 Relief 算法进行了进一步研究，改善了 Relief 只能适用于二分类的情况，改进后的算法能针对多类任务。Dash 和 Liu 在 1997 年提出现在有的特征提取算法都认为特征之间是没有交互信息，对存在的特征选择算法进行了总结[8]。他们对搜索的方法和评价准则进行了严格的分类。其中根据搜索方法的不同可以分为：完全搜索方法、随机搜索方法、启发式搜索方法等。

我国的学者陈彬在 1990 后，提出了最优特征子集是 NP-hard 问题，对于之后的研究有非常关键的推进作用[9]。从今天来看，随着大数据时代到来，在数据量暴涨的今天，特征数量非常庞大，即使是良好性能的计算机对于完成搜索问题都是很有难度的。

LuisC M 等人在 2002 年对现存的特征选择算法进行完整的回顾，并对这些算法进一步的评价和比较。D.A.Clausi 等人提出了 KIF 方法，使用了无监督聚类的 Fisher 准则[10]。

近年来，不同的特征选择算法用不同评价准则，加入新搜索算法，进一步改进了算法效率。比如监督学习算法，半监督学习算法，马尔科夫算法[11]，支持向量机，神经网络算法，遗传算法，信息熵法。同时，Filter 类算法、Wrapper 类算法、Embedded 类算法也得到了很大的重视和发展。